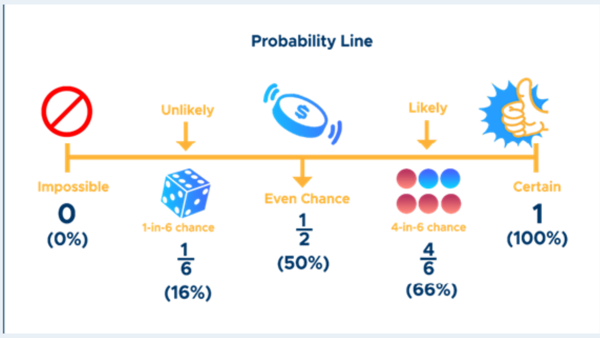
Probability & Variability Cheat Sheet

Hamed Ahmadinia, Ph.D.

Hamed.Ahmadinia@metropolia.fi

Probability Basics

What is Probability?
Probability tells you how likely something is to happen, from 0 (impossible) to 1 (certain).



Explanation of the Image:
This line shows how chance moves from 0% (impossible) on the left to 100% (certain) on the right.

- A 1-in-6 chance (about 16%) is like rolling a specific number on a fair die.
- A 1-in-2 chance (50%) is like flipping a fair coin.
- A 4-in-6 chance (66%) means it is more likely than not.

Key Points:

- All possible outcomes add up to 1 (or 100%).
- **Discrete:** Countable outcomes (like dice faces).
- **Continuous:** Any value in a range (like time).

Expected Value

What is Expected Value?
It is like the average result if you do the same experiment many times.

Formulas:

- **Discrete:** $E(X) = \sum x\, P(x)$
- **Continuous:** $E(X) = \int x\, f(x)\, dx$

Example:
For a fair dice, the expected value is:

$$E(X) = \frac{1+2+3+4+5+6}{6} = 3.5$$

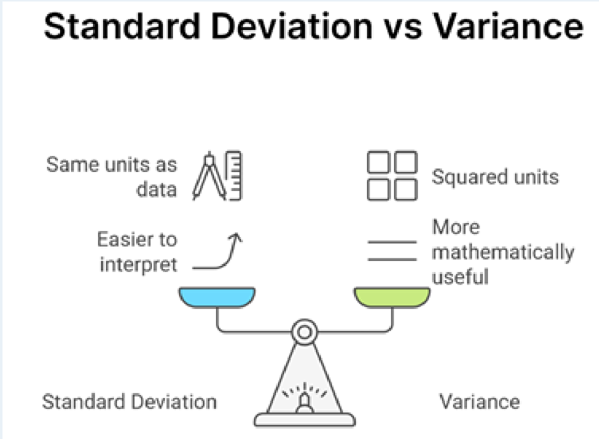
C8 | =SUM(C2:C7) | EV = ΣPx * X

	A	B	C	D
1	Dice (X)	Probability [P(X)]	X * P(X)	
2	6	0.166666667	1	
3	5	0.166666667	0.83	
4	4	0.166666667	0.67	
5	3	0.166666667	0.50	
6	2	0.166666667	0.33	
7	1	0.166666667	0.17	
8	Expected Value		3.50	

Simple Idea:
Multiply each outcome by its probability and sum these products to get the long-term average.

Variance & Standard Deviation

What do they show?
They show how spread out the data is.



Explanation of the Image:
- Standard Deviation uses the same units as your data, making it easier to explain. - Variance uses squared units, which is more useful for certain math but harder to interpret directly.

Definitions:

- **Variance:** Average of the squared differences from the mean.
- **Standard Deviation:** The square root of the variance.

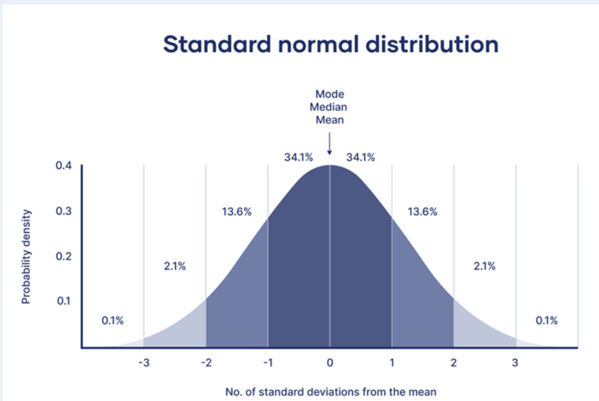
$$\mathrm{Var}(X) = E[(X-\mu)^2], \quad \sigma = \sqrt{\mathrm{Var}(X)}$$

Example:
If most test scores are close to the average, the standard deviation is small. If they vary a lot, it's large.

Simple Idea:
They help you see if your data is packed together or spread out.

Normal Distribution

What is the Normal Distribution?
A bell-shaped curve where most values are near the mean and fewer are far away.



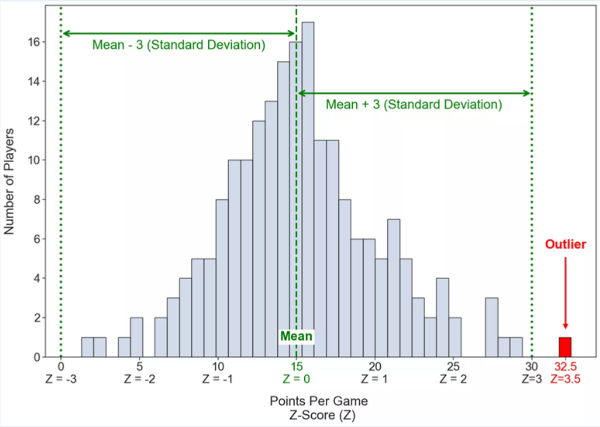
Key Points:

- Mean μ: center of the curve.
- Variance σ^2: controls how wide it is.
- Symmetrical (same shape on both sides).

Z-Scores & Outliers

What is a Z-Score?
A z-score tells you how many standard deviations a value is away from the mean:

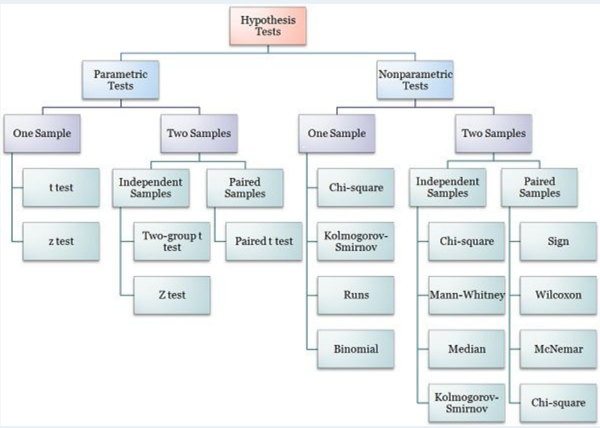
$$z = \frac{x - \mu}{\sigma}$$



Explanation of the Image:
- The histogram shows points per game for 200 basketball players. - The center (Z=0) is the mean. - Values beyond 3 standard deviations (e.g., Z=3 or higher) are often considered outliers.

Statistical Tests

What are Statistical Tests?
They check if differences in data are real or just random.



Explanation of the Image:
- Shows different types of hypothesis tests, both parametric (assuming normality) and nonparametric (no distribution assumption). - Each branch helps decide which test to use based on how many samples you have and whether your data is paired or independent.

Types:

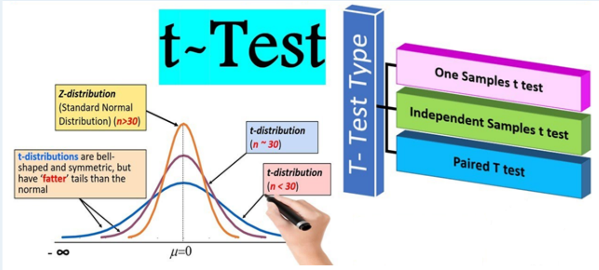
- **Parametric (T-test, Z-test):** Assumes normal data.
- **Non-Parametric (Mann-Whitney, Chi-Squared, etc.):** No normal assumption.

Example:
A t-test can see if Class A's average score is truly higher than Class B's average or if it's just luck.

Simple Idea:
They help decide if an effect or difference is big enough to matter.

T-Test (Parametric)

When to Use:
- Comparing the means of two groups (e.g., two classes). - Data is (roughly) normally distributed or sample size is moderate/large.

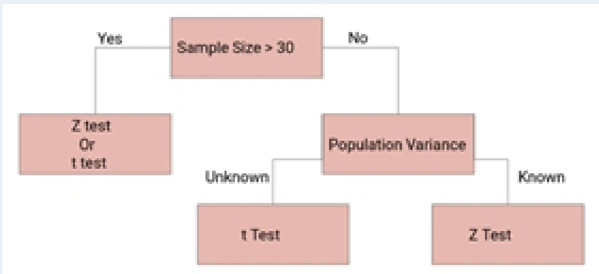


Explanation of the Image:
- The t-Test compares sample means and determines if the difference is significant. - If $n > 30$, the t-distribution approximates the normal (Z-distribution). - If $n < 30$, the t-distribution has "fatter" tails than the normal.
Example:

```
from scipy import stats

group1 = [70, 72, 68, 74, 71]
group2 = [65, 67, 64, 66, 68]
tstat, pval = stats.ttest_ind(group1, group2)

if pval < 0.05:
    print("Significant difference")
else:
    print("No significant difference")
```

Z-Test (Parametric)

When to Use:
- Large sample size ($n > 30$) or known population variance. - Compare sample mean to a known/hypothesized mean.



Explanation of the Image:
- If the sample size is **greater than 30**, you can use either a **Z-test** or **t-test**. - If the sample size is **less than 30**, check whether the **population variance is known**:

- If the variance is **known**, use a **Z-Test**.
- If the variance is **unknown**, use a **t-Test**.

Example:

```
# pip install statsmodels
from statsmodels.stats.weightstats import ztest

data = [68, 70, 72, 71, 69, 74, 67]
z_stat, p_val = ztest(data, value=70)

if p_val < 0.05:
    print("Reject H0 (mean != 70)")
else:
    print("Fail to reject H0")
```

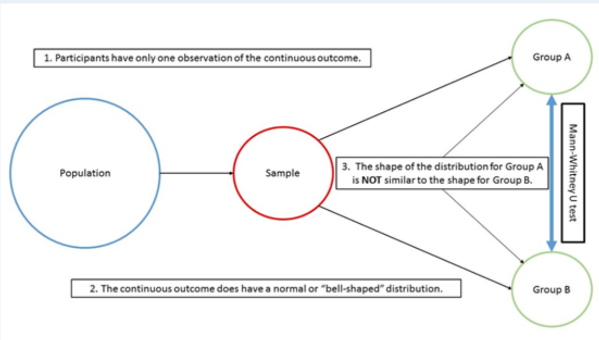
Chi-Squared Test (Non-Parametric)

When to Use:
- Categorical data (frequencies in categories). - Compare observed counts to expected counts.
Example:

```
from scipy import stats

observed = [50, 30, 20]  # observed frequencies
expected = [40, 40, 20]  # expected frequencies

chi2, p_val = stats.chisquare(f_obs=observed, f_exp=expected)

if p_val < 0.05:
    print("Reject H0 (distribution differs)")
else:
    print("Fail to reject H0")
```

Mann-Whitney U (Non-Parametric)

When to Use:
- Two independent groups. - Data not necessarily normal; uses ranks instead of means.



Explanation:
- The Mann-Whitney U test is used when:

- Participants have only **one observation** of the continuous outcome.
- The continuous outcome has a **normal or bell-shaped distribution**.
- The distribution shape of **Group A is NOT similar** to the shape of **Group B**.

Example:

```
from scipy import stats

groupA = [12, 15, 14, 10, 9]
groupB = [18, 17, 20, 19, 22]

u_stat, p_val = stats.mannwhitneyu(groupA, groupB,
alternative='two-sided')

if p_val < 0.05:
    print("Reject H0 (distributions differ)")
else:
    print("Fail to reject H0")
```

Kolmogorov-Smirnov (Non-Parametric)

When to Use:
- Compare two distributions (or a sample vs. a known distribution). - Checks if two samples come from the same distribution.
Example:

```
from scipy import stats

sample1 = [1.2, 1.4, 1.6, 1.8, 2.0]
sample2 = [1.3, 1.5, 1.5, 1.9, 2.1]

ks_stat, p_val = stats.ks_2samp(sample1, sample2)

if p_val < 0.05:
    print("Reject H0 (distributions differ)")
else:
    print("Fail to reject H0")
```

Wilcoxon Test (Non-Parametric)

When to Use:
- Paired data (e.g., before/after) with non-normal distribution.
- Similar to paired T-test but for non-parametric data.
Example:

```
from scipy import stats

before = [5, 7, 6, 8, 7]
after  = [6, 9, 7, 9, 8]

w_stat, p_val = stats.wilcoxon(before, after)

if p_val < 0.05:
    print("Reject H0 (median difference != 0)")
else:
    print("Fail to reject H0")
```

Website: ahmadinia.fi