HUMBEHV 3ST3

one-sample t-tests

Prof. Patrick Bennett

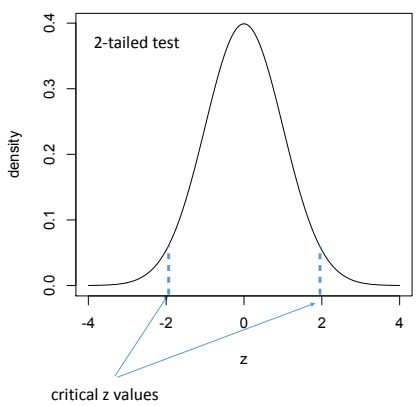
Testing Hypotheses with Known Population Variance

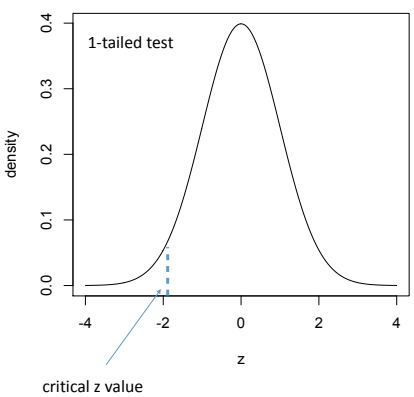
Tapping test (single observation)

- tapping test from Chapter 8
- tapping rate can be used as an indicator for neurological damage
- Chistianson & Leathem (2004)
  - 10 s tapping test in normal pop: mean = 59; sd = 7
  - tapping rate is slower in Alzheimer's patients (& other clinical populations)
  - **measured rate for 1 subject = 45 (per 10 s trial)**
- Question: Is this subject drawn from typical or Alzheimer population?
- Will use null hypothesis significance testing to answer this question…
  - **What null (H0) & alternative (H1) hypotheses should we use?**

General Strategy

reject H0 if z exceeds **critical values** of z





Tapping test (single observation)

$$z = \frac{\bar{X} - \mu}{\sigma}$$

- measured rate for 1 case = 45 (per 10 s trial)
- is this case unusual?
  - H0: μ ≥ 59 (case was <u>not</u> drawn from Alzheimer's population)
  - H1: μ < 59 (case was drawn Alzheimer's population)
- significance level of .05 (one-tailed)
  - critical value of z = -1.645; p(z ≤ -1.645 | H0) = .05
- observed z = (45-59)/7 = -2.0
- observed z < critical z; p(z ≤ -2 | H0) = 0.023
  - reject H0 in favour of H1

Tapping test (group mean)

- Instead of single case study, consider a situation in which we measure the mean tapping rate of 5 individuals
  - administer genetic test to screen for early-onset Alzheimer's disease to many individuals
  - identify 5 individuals who might be at risk
  - also administer tapping test to these individuals
- Question: is tapping rate for this <u>group</u> unusually low?
  - N=5; mean = 54.2; standard deviation = 6

Tapping test (group mean)

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

- Population: μ = 59, σ = 7
- Sample (N=5): mean = 54.2; standard deviation = 6
- Question: is tapping rate for this <u>group</u> unusually low?
  - H0: μ ≥ 59 (sample was not drawn from Alzheimer's population)
  - H1: μ < 59 (sample was drawn Alzheimer's population)
- Sampling distribution **<u>of mean</u>** assuming H0 is true:
  - $N(\mu,\sigma^2) = N(59, 7^2/5) = N(\mu=59,\sigma^2=9.8)$ [σ = sqrt(9.8) = 3.13]
- z score for our mean: z = (54.2-59) / 3.13 = -1.533
- significance level = .05 (one-tailed); Critical z = -1.645; p(z ≤ -1.645 | H0) = 0.05
- Observed z (-1.533) > Critical z (-1.645)
  - p(z ≤ -1.533 | H0) = 0.063
  - fail to reject H0

Testing Hypotheses with Unknown Population Variance

Tapping test (group mean)

- Population: μ = 59, **σ = ?**
- Sample (N=5): mean = 54.2; standard deviation = 6
- Question: is tapping rate for this group unusually low?
  - H0: μ ≥ 59 (sample was not drawn from Alzheimer's population)
  - H1: μ < 59 (sample was drawn Alzheimer's population)
- Sampling distribution **of mean** assuming H0 is true:
  - $N(\mu,\sigma^2) = N(59, 6^2/5) = N(\mu=59,\sigma^2=7.2)$ [σ = sqrt(7.2) = 2.68]
- z score for our mean: "z" = (54.2-59) / 2.68 = -1.79
  - N.B. This "z" is based on ESTIMATED standard deviation
- significance level = .05 (one-tailed); Critical z = -1.645; p(z ≤ -1.645 | H0) = 0.05
- Estimated "z" (-1.79) < Critical z (-1.645)
  - p(z ≤ -1.79 | H0) = 0.037
  - reject H0 in favour of H1

Effect of using estimate of σ

- z is defined with KNOWN population μ and σ
- only source of variation in z is sampling error of mean
- using estimate of σ introduces another source of variation in z
  - Estimated z depends on group mean AND sample standard deviation
- how does this affect distribution of z?

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

$$\hat{z} = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

Effect of using estimate of population variation

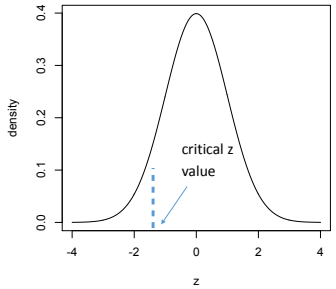
- William Gossett applied statistics to his work in the Guinness brewery
- Under the pseudonym, Student, he investigated effects of estimating σ on z test
  - sample variance is unbiased estimate of population variance
  - but sample standard deviation is a **biased** estimate of population standard deviation
  - sample SD underestimates population SD particularly for small sample sizes
- Discovered that **using estimates of σ lead to more extreme values of z than predicted** by statistical theory

William Sealy Gosset
(aka Student)

Effect of inflating z score

$$\hat{z} = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- calculating z with estimated σ inflates z scores
- extreme $\hat{z}$ values **occur more frequently** than expected when H0 is True
- **what effect does this have on our evaluation of H0?**



Critical z = -1.645

p(z ≤ -1.645 | H0) = 0.05

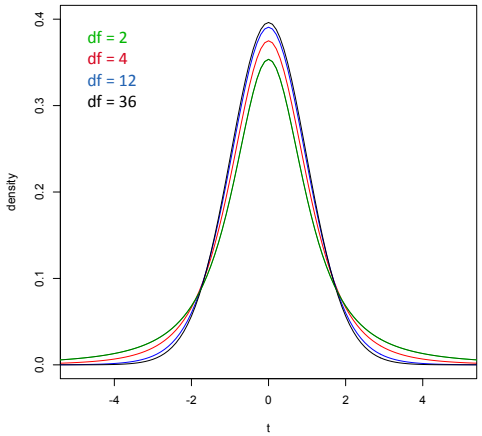
Effect of using estimate of population variation

- William Gossett applied statistics to his work in the Guinness brewery
- Under the pseudonym, Student, he investigated effects of estimating σ on z test
- Discovered that using estimates of σ lead to more "extreme" values of z than predicted by statistical theory
- Caused an increase in Type I errors
  - especially for small samples
- Devised a new test that corrected these errors
  - Student's *t* test

William Sealy Gosset
(aka Student)

t distribution

- unimodal
- symmetrical around zero
- has 1 parameter:
  - degrees of freedom (df)
- df alters kurtosis
  - lower df associated with narrower middle portion & heavier tails
- t approximately normal for df ≅ 35



what are "degrees-of-freedom"?

degrees of freedom

- consider a set of n=4 numbers:
  - guess the value of each number:
    - 2
    - 0
    - 8
    - 10
- hard to guess correctly because each number can be <u>any</u> value
  - each number is free to vary

degrees of freedom

- consider a set of n=4 numbers, whose total = 20
  - guess the value of each number:
    - 4
    - 1
    - 10
    - 5
- first 3 numbers can be <u>any</u> value
  - but value of 4th is determined by first 3 (and the total value of 20)
  - 4th value is <u>not</u> free to vary
- given the total, we say the set of n=4 values has n-1=3 degrees of freedom

degrees of freedom

- consider a set of n=4 numbers, whose mean = 5
  - guess the value of each number:
    - 5
    - 8
    - 6
    - 1
- first 3 numbers can be <u>any</u> value
  - but value of 4th is determined by first 3 (and the mean value of 5)
  - 4th value is <u>not</u> free to vary
- given the mean, we say the set of n=4 values has n-1=3 degrees of freedom

degrees of freedom

Essentially the term means the number of independent units of information in a sample relevant to the estimation of a parameter or calculation of a statistic.

— B.S. Everitt, The Cambridge Dictionary of Statistics

Degrees-of-freedom (df)

- Degrees of freedom show up in <u>many</u> different places in statistics
- when calculating *t* for 1 sample
  - df = sample size minus one = n-1
  - true because *t* is based on sample variance
  - which depends on sample <u>mean</u>
  - given the mean, only (n-1) sample values are free to vary
    - the nth-value is determined by the other n-1 values

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

Back to hypothesis testing

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

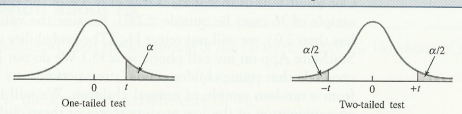
- When σ is NOT known
  - estimated z is inflated
  - our standardized score does <u>not</u> follow z distribution
  - using "z" increases Type I error rate
- However, standardized score <u>DOES</u> follow a *t* distribution
- Therefore, our estimated "z" actually is a t statistic
- and we use critical values of *t*, not *z*, to evaluate null hypothesis

Back to hypothesis testing

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- Our tapping test sample: n=5, df=n-1=4
  - significance level = .05 (one-tailed)
  - **<u>critical t(df=4) = -2.13</u>**
    - p(t ≤ -2.13 | H0) = .05

Table 12.1 Abbreviated Version of Table E.6, Percentage Points of the *t* Distribution



	Level of Significance for One-Tailed Test								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test								
df	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	63.662
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
...	...	...	...	...	...	...	...	...	...
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	.679	.849	1.047	1.299	1.676	**2.009**	2.403	2.678	3.496
100	.677	.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Source: The entries in this table were computed by the author.

Back to hypothesis testing

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- Our tapping test sample: n=5, df=n-1=4
  - significance level = .05 (1-tailed); <u>critical t(df=4) = -2.13</u>
  - p(t ≤ -2.13 | H0) = .05
- Tapping example using z (reminder of key results):
  - significance level = .05 (1-tailed); Critical z = -1.645; p(z ≤ -1.645 | H0) = 0.05
  - Observed Estimated "z" (-1.79) < Critical z (-1.645); reject H0 in favour of H1
  - p(z ≤ -1.79 | H0) = 0.037; reject H0 in favour of H1

Back to hypothesis testing

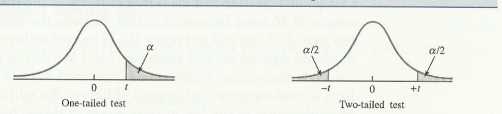
$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- Our tapping test sample: n=5, df=n-1=4
  - significance level = .05 (1-tailed); critical t(df=4) = -2.13
  - p(t ≤ -2.13 | H0) = .05
- Tapping example using z (reminder of key results):
  - significance level = .05 (1-tailed); Critical z = -1.645; p(z ≤ -1.645 | H0) = 0.05
  - Observed Estimated "z" (-1.79) < Critical z (-1.645); reject H0 in favour of H1
  - p(z ≤ -1.79 | H0) = 0.037; reject H0 in favour of H1
- Tapping test example using t:
  - **Observed t** = (54.2-59) / 2.68 = -1.79; p(t ≤ -1.79 | H0) = 0.074
  - Observed t (-1.79) not less than Critical t (-2.13)
  - so, do NOT reject H0 in favour of H1

two-sided t test (tapping test example)

- Null & Research Hypotheses:
  - H0: μ=59 (sample drawn from healthy population)
  - H1: μ≠59 (sample not drawn from healthy population)
- alpha = .05
  - critical values of t = ± 2.776 [df=4, 2-tailed, alpha=.05]

Table 12.1 Abbreviated Version of Table E.6, Percentage Points of the *t* Distribution



	Level of Significance for One-Tailed Test								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Level of Significance for Two-Tailed Test								
df	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	63.662
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
...	...	...	...	...	...	...	...	...	...
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	.679	.849	1.047	1.299	1.676	**2.009**	2.403	2.678	3.496
100	.677	.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Source: The entries in this table were computed by the author.

two-sided t test (tapping test example)

- Null & Research Hypotheses:
  - H0: μ=59 (sample drawn from healthy population)
  - H1: μ≠59 (sample not drawn from healthy population)
- alpha = .05
  - critical values of t = ± 2.776 (df=4, 2-tailed, alpha=.05)
  - given H0, p(t > 2.776) = .025 & p(t < -2.776) = .025
- sample: N=5, mean = 54.2, standard deviation = 6
  - following values are the same as for 1-tailed tests:
  - t = (54.2-59) / sqrt(6^2/5) = (54.2-59)/2.68 = –1.79
- Observed t (-1.79) is not more extreme than either critical t value (±2.776)
  - fail to reject H0

effect of sample size (tapping test example)

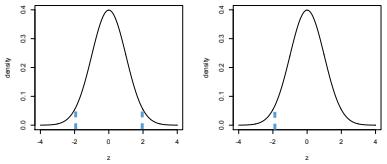
$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}}$$

- Null & Research Hypotheses:
  - H0: μ=59 (sample drawn from healthy population)
  - H1: μ≠59 (sample not drawn from healthy population)
- sample: **N=30**, mean = 54.2, standard deviation = 6
  - t = (54.2-59) / **sqrt(6^2/30)** = (54.2-59)/**1.09** = **–4.4**
  - SD of sampling distribution decreases from sqrt(6^2/5) to sqrt(6^2/30)
    - so same values of μ, sample mean, and sd result in bigger t value
- alpha = .05 (df=n-1=29)
  - critical values of t = ± 2.04 [approximate; value for df=30 taken from Table 12.1]
  - given H0: p(t < -2.04) = .025 & p(t > 2.04) = .025
  - p({t < -2.04 **OR** t > 2.04} | H0) = .025 + .025 = .05
- Observed t (–4.0) is more extreme than either critical t value (±2.04)
  - reject H0 in favour of H1

Factors affecting *t* test

$$t = \frac{\bar{X} - \mu_{\bar{X}}}{\hat{\sigma}_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{s/\sqrt{N}}$$

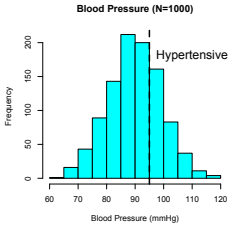
- observed difference between sample mean - μ (when H0 is true)
- sample standard deviation (s)
- sample size (N)
  - p-value depends on N
  - large N means smaller p values
- significance level (α)
- 1- vs. 2-tailed test

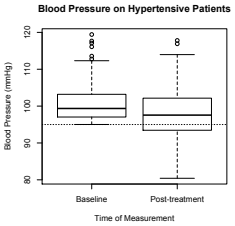


Paired-sample t test (regression to the mean)

Treatment for Hypertension (High Blood Pressure)

- measure blood pressure (BP) in 1000 adults
- select 296 adults HBP > 95 mmHG
  - apply treatment
  - measure post-treatment BP
- compare pre- & post-treatment BP
  - t-test on paired samples
  - each subject gives 2 measures
  - do t test on **difference** scores

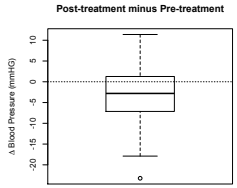


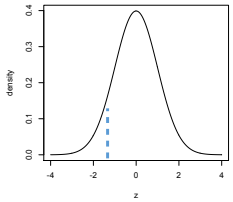


Treatment for Hypertension (High Blood Pressure)

- Difference: mean = -2.91, sd = 6.02, N = 296
  - H0: True Difference >= 0
  - H1: True Difference < 0
- critical value of t:
  - df = N-1 = 295 & alpha = .05
  - t.critical = -1.65
  - p(t≤ -1.65 | H0) = 0.05
- observed value of t:
  - t= (-2.91 - 0)/sqrt(6.02^2/296) = -8.31
- p(t ≤ -8.31 | H0) < 0.0001
  - reject H0 in favour of H1

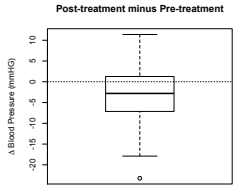
median = -2.81
IQR = [1.26, -7.1] = 8.36





Treatment for Hypertension (High Blood Pressure)

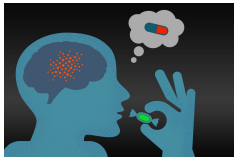
- Difference: mean = -2.91, sd = 6.02, N = 296
  - H0: True Difference = 0
  - H1: True Difference < 0
- Reject H0 in favour of H1
- Does that mean our "treatment" worked?
  - Did the treatment cause the reduction in BP?
- Not necessarily!



Is reduction in HBP due to <u>treatment</u>?

What else might cause blood pressure to be lower in the 2nd (post-treatment) test?

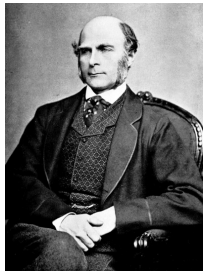
- Placebo Effect?
- Increased familiarity, reduced anxiety on 2nd test
- How could we rule out these alternative explanations?
- One more possible explanation:
  - Regression to the Mean



The weird power of the placebo effect, explained
Brian Resnick, Vox, 2017

Regression to the Mean

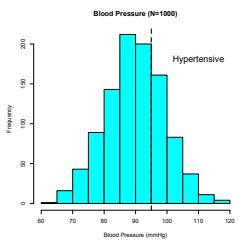
- Very tall parents tend to have shorter children
  - & very short parents tend to have taller children
- Students with very high scores on test 1 tend to have slightly lower scores on test 2
  - & students with very low scores on test 1 tend to have slightly higher scores on test 2
- extreme scores on 1st measurement tend to be closer to average on 2nd measurement

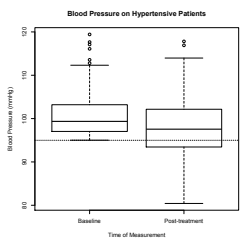


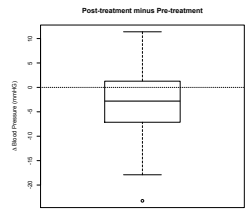
Francis Galton

Regression to the Mean

Note that blood pressure experiment included only subjects with high BP



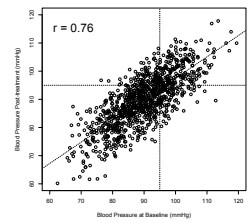




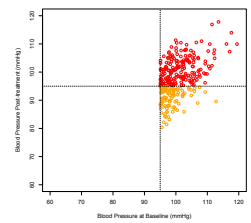
Regression to the Mean

Note that blood pressure experiment included only subjects with high BP

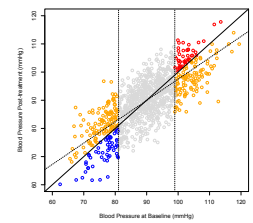
Pre & Post treatment scores for all subjects



Pre & Post treatment scores for high BP subjects

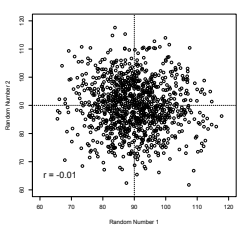


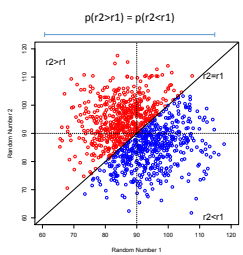
Pre & Post treatment scores for low BP subjects

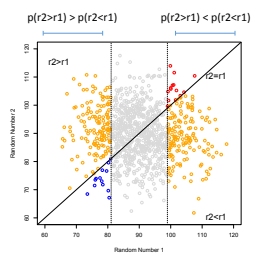


Regression to the Mean

Regression to the mean reflects probabilistic variation across measurements





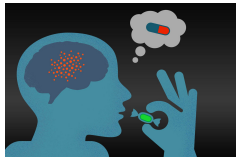


Regression to the mean occurs even for totally random data!

Is reduction in HBP due to <u>treatment</u>?

What else might cause blood pressure to be lower in the 2nd (post-treatment) test?

- Placebo Effect?
- Increased familiarity, reduced anxiety on 2nd test
- How could we rule out these alternative explanations?
- One more possible explanation:
  - Regression to the Mean
  - Selecting extreme high scores on test 1 means that we should expect a lower average on test 2 <u>even if the drug/treatment has no effect!</u>
  - one way to test this idea: include a no-treatment group



The weird power of the placebo effect, explained
Brian Resnick, Vox, 2017

1-sample t tests (summary)

- z tests used when population variance is known
- estimating population variance from sample variance inflates “z”
  - estimated-z is not distributed as standardized normal variable
    - extreme values occur more frequently than expected
  - “z” test has higher Type I error rate than expected
- t tests used when population variance is estimated from sample
  - logic of t test is the same as z test
  - primary difference is we compare observed t to critical value of t, not z
  - corrects for inflation of Type I error rate
  - t test on paired samples is a t test on difference scores
- Be cautious when interpreting results!
  - unusually low or high scores/means, or a difference between pre- & post-treatment scores, may occur for several (sometimes non-obvious) reasons!