Stat 431 Midterm Review Guide, Hyunseung Kang

All of Stat 431 starts by (i) **collecting i.i.d. samples**, $X_1, \dots, X_n$ from a distribution and (ii) **studying its properties** of the distribution. The main focus is studying its **mean**,μ, and **variance**, σ^2.

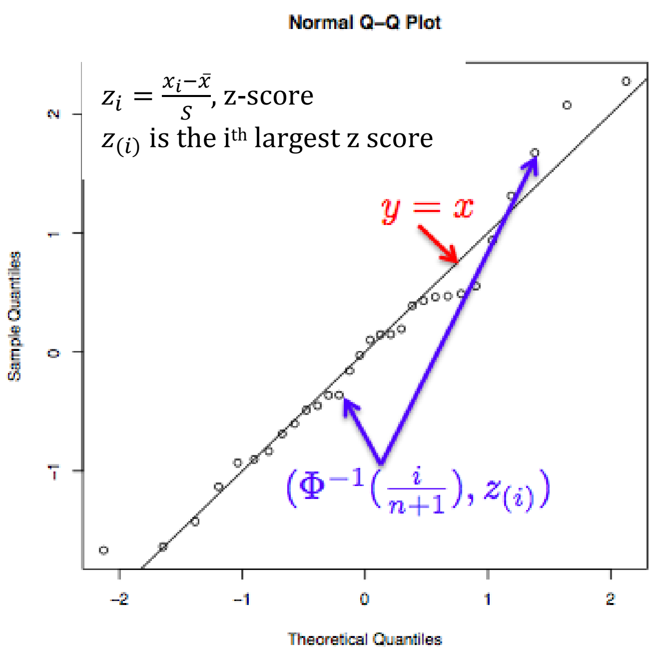
1. Sampling Distribution

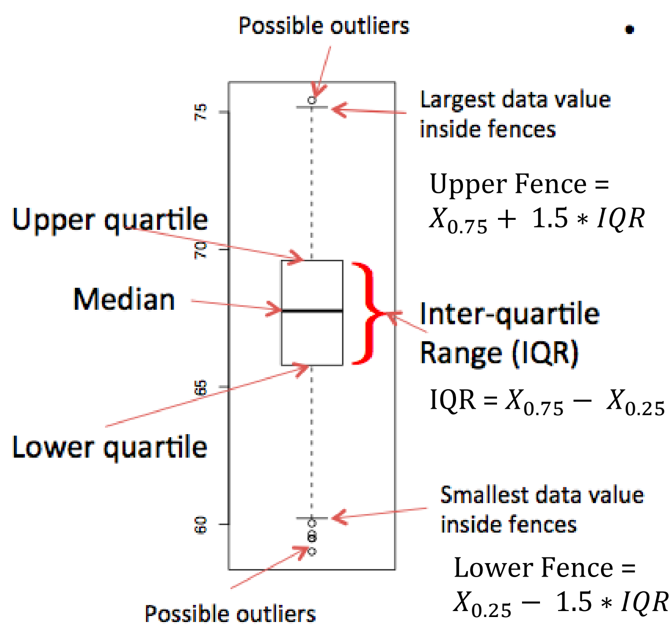
The most natural thing to do once we collected our sample is to compute its sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and sample variance, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})$. In addition, we can identify the distributions of them

	Normal Distribution	t Distribution	Chi-Square Distribution
Parameters of Interest (for testing and CI)	μ with known σ	μ with unknown σ	σ
Examples	1. $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ 2. $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ 3. $Z \sim N(0,1)$	1. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ 2. $\frac{Z}{\sqrt{\frac{U}{n}}} \sim t_n$ where $Z \sim N(0,1)$, $U \sim \chi^2_n$, and U and Z are indep.	1. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ 2. $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$, where $Z_i \sim N(0,1)$ and are i.i.d.
Properties	1. If $n \geq 30$, then regardless of the original distribution of X_i, $\bar{X}$ is normally distributed by Central Limit Theorem. Otherwise, $X_i \sim N(\mu, \sigma^2)$ assumption is needed	1. If df (i.e. $n-1$) gets larger, (i.e. $df \geq 30$) then $t \sim N(0,1)$ 2. t Distribution has a "FATTER" tail than the normal distribution → the 95% intervals using t are LARGER than those using normal	

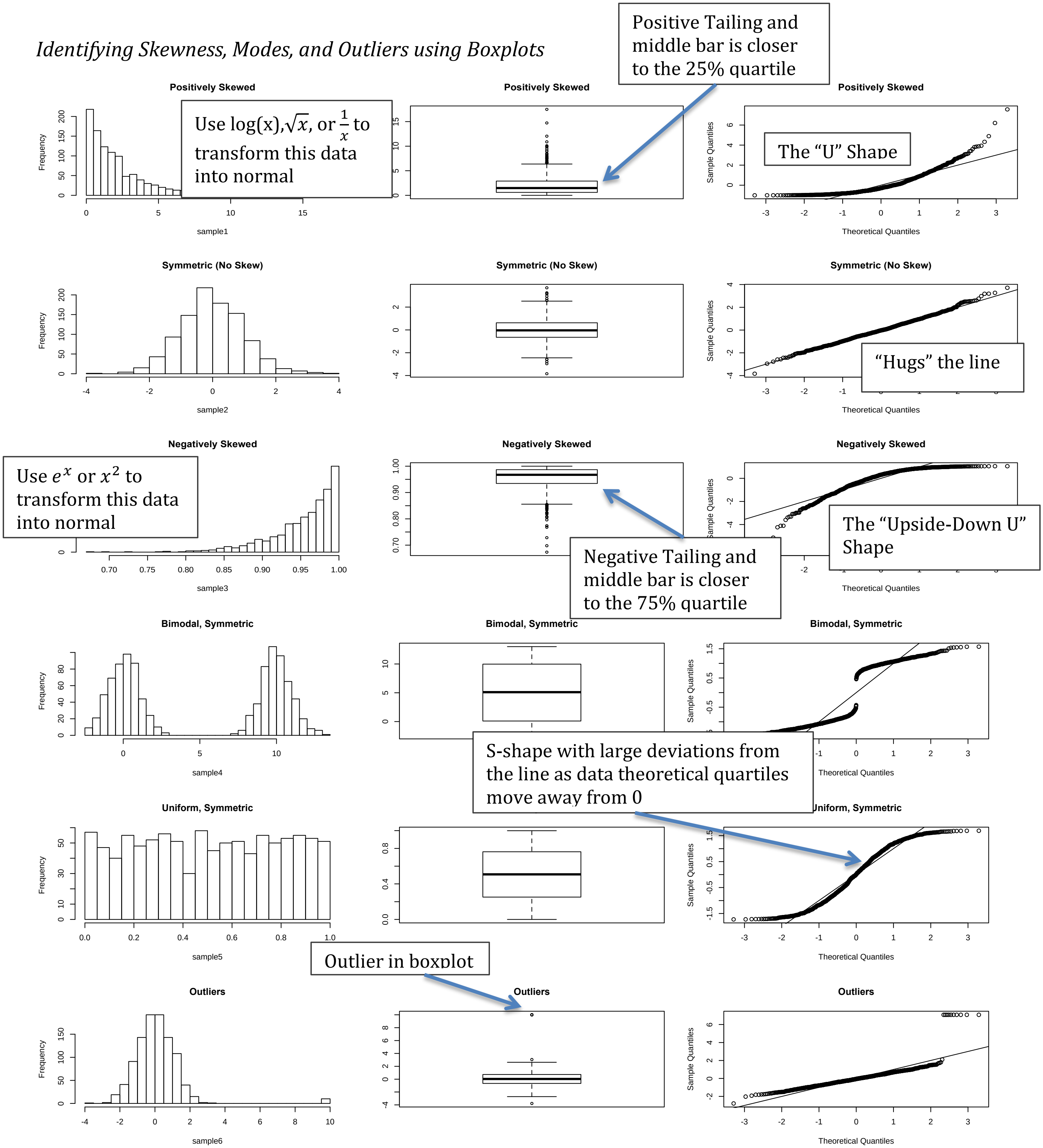
Also, we can use boxplots/qq plots to summarize the sampling distribution

Basics of QQ plots and Boxplots





Identifying Skewness, Modes, and Outliers using Boxplots



2. Testing Mean and Variance for One Sample

In addition to using summary statistics and graphical tools, we can hypothesize about what the true μ or σ may entail and test our hypothesis. In particular, the procedure of hypothesis testing is as follows:

1. We hypothesis whether it's a **two-sided** or a **one-sided test**.
2. We create a **test statistic, T(X),** to test (a). All of our test statistics have to be **pivotal** (i.e. the probabilities of our test statistics CANNOT depend on the true unknown parameters).

3. We test our hypothesis with the test statistic. In particular, a natural testing procedure would be to **reject the null is the probability of rejecting it is small** given that the null is indeed true (i.e. Type I error is small). For example, for the two-sided case we reject the null whenever $P_{\mu_0}(|T(X)| > t(x)) < \alpha$ where $t(x)$ is the observed value of the test statistic and **α is the maximum probability of Type I Error.**

The truth / Our conclusion from data	H_0 is actually true	H_a is actually true **Note**: Type II + Power = 1
We conclude: H_0 is correct (i.e. $H_0$is retained)	**Note**: We don't care about this box since we always control for Type I Error (aka this term is also controlled by $1-\alpha$)	**Type II Error** or $P(retain\ null \mid H_a is\ true)$
We conclude: H_a is correct (i.e. H_0 is rejected)	**Type I Error** or $P(reject\ null \mid H_0 is\ true)$ **Note**: always controlled by α	**Power** or $P(reject\ null \mid H_a is\ true)$

We can also reject the null through a variety of methods, as listed below.

a. **P-values**: If the maximum probability of Type I error from the sample (i.e. p-value) is less than the allowed probability of Type I Error (i.e. $P\left(|Z| > \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) < \alpha$), we reject the null

b. **Rejection Region**: We reject the null whenever our observed test statistic is in a spec. region.

c. **Confidence Intervals**: If the interval DOES NOT contains μ_0, we REJECT the null. Note that confidence intervals are random and it covers μ under repeated sampling $1-\alpha$ amount of times.

	Testing μ with known σ^2	Testing μ with unknown σ^2	Testing σ^2
Hypothesis Setup	1. $H_0: \mu = \mu_0, H_a: \mu \neq \mu_0$ (two-sided) 2. $H_0: \mu \leq \mu_0, H_a: \mu > \mu_0$ (one-sided) 3. $H_0: \mu \geq \mu_0, H_a: \mu < \mu_0$ (one-sided) **Note**: The direction of the inequality sign in the rejection region is **identical** to the direction of the inequality sign in the alternative		1. $H_0: \sigma = \sigma_0, H_a: \sigma \neq \sigma_0$ 2. $H_0: \sigma \leq \sigma_0, H_a: \sigma > \sigma_0$ 3. $H_0: \sigma \geq \sigma_0, H_a: \sigma < \sigma_0$
Assumptions	i.i.d. sample from a normal distribution*		
Test Statistic	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$	$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$
Rejection Region	1. $|\bar{x}-\mu_0| > z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ 2. $\bar{x} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ 3. $\bar{x} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$	1. $|\bar{x}-\mu_0| > t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$ 2. $\bar{x} > \mu_0 + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$ 3. $\bar{x} < \mu_0 - t_{\alpha,n-1}\frac{s}{\sqrt{n}}$	1. $S^2 > \frac{\sigma_0^2}{n-1}\chi^2_{\frac{\alpha}{2},n-1}$ or $S^2 < \frac{\sigma_0^2}{n-1}\chi^2_{1-\frac{\alpha}{2},n-1}$ 2. $S^2 > \frac{\sigma_0^2}{n-1}\chi^2_{\alpha,n-1}$ 3. $S^2 < \frac{\sigma_0^2}{n-1}\chi^2_{1-\alpha,n-1}$
Confidence Intervals	1. $(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$ 2. $(\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty)$ 3. $(-\infty, \bar{x} + z_\alpha\frac{\sigma}{\sqrt{n}})$	1. $(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}})$ 2. $(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \infty)$ 3. $(-\infty, \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}})$	1. $\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)$ 2. $\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, \infty\right)$ 3. $\left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$
Power: **NOTE**: Need a pre-specified	1. $P_{\mu \in H_a}\left(\frac{|\bar{X}-\mu_0|}{\sigma/\sqrt{n}} > z_{\frac{\alpha}{2}}\right)$ 2. $P_{\mu \in H_a}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha\right)$	1. $P_{\mu \in H_a}\left(\frac{|\bar{X}-\mu_0|}{S/\sqrt{n}} > t_{\frac{\alpha}{2},n-1}\right)$ 2. $P_{\mu \in H_a}\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{\alpha,n-1}\right)$	1. $P_{\sigma^2 \in H_a}\left(\frac{S^2(n-1)}{\sigma_0^2} > \chi^2_{\frac{\alpha}{2},n-1} \cup \frac{S^2(n-1)}{\sigma_0^2} < \chi^2_{1-\frac{\alpha}{2},n-1}\right)$

value from the alternative	3. $P_{\mu \in H_a}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < z_\alpha\right)$	3. $P_{\mu \in H_a}\left(\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < t_{\alpha,n-1}\right)$	2. $P_{\sigma^2 \in H_a}\left(\frac{S^2(n-1)}{\sigma_0^2} > \chi^2_{\alpha,n-1}\right)$ 3. $P_{\sigma^2 \in H_a}\left(\frac{S^2(n-1)}{\sigma_0^2} < \chi^2_{1-\alpha,n-1}\right)$

*You can check for normality using the QQ plot and make transformations as needed. Also, if the sample size is greater than 30, then by the CLT, we can assume it came from a normal distribution

ou

3. Testing Mean for Two Samples

Suppose we collect $X_1, \dots, X_{n_1}$ i.i.d. from sample 1 and $Y_1, \dots, Y_{n_2}$ i.i.d. from sample 2. We want to test whether the two samples have the same means (i.e. $\mu_1 = \mu_2$)

	Equal Variance Assumption	Without Equal Variance Assumption
Hypothesis Setup	1. $H_0: \mu_1 - \mu_2 = \mu_0, H_a: \mu_1 - \mu_2 \neq \mu_0$ (two-sided) 2. $H_0: \mu_1 - \mu_2 \leq \mu_0, H_a: \mu_1 - \mu_2 > \mu_0$ (one-sided) 3. $H_0: \mu_1 - \mu_2 \geq \mu_0, H_a: \mu_1 - \mu_2 < \mu_0$ (one-sided) **Note**: The direction of the inequality sign in the rejection region is **identical** to the direction of the inequality sign in the alternative	
Assumptions	i.i.d samples from two normal distribution* AND independent samples from each other	
Test Statistic	A. $\frac{(\bar{X}-\bar{Y})-\mu_0}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$, we don't know σ^2 B. $\frac{(\bar{X}-\bar{Y})-\mu_0}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$, we know σ^2** $S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$	A. $\frac{(\bar{X}-\bar{Y})-\mu_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_v$, we don't know σ^2 B. $\frac{(\bar{X}-\bar{Y})-\mu_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$, we know σ^2** $\min(n_1,n_2) - 1 \leq v \leq n_1+n_2-2$
Rejection Region	1. $\lvert\bar{x}-\bar{y}-\mu_0\rvert > t_{n_1+n_2-2,\frac{\alpha}{2}} S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ 2. $\bar{x}-\bar{y} > \mu_0 + t_{n_1+n_2-2,\alpha} S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ 3. $\bar{x}-\bar{y} < \mu_0 - t_{n_1+n_2-2,\alpha} S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	1. $\lvert\bar{x}-\bar{y}-\mu_0\rvert > t_{v,\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ 2. $\bar{x}-\bar{y} > \mu_0 + t_{v,\alpha}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ 3. $\bar{x}-\bar{y} < \mu_0 - t_{v,\alpha}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$
Confidence Intervals	1. $\bar{x}-\bar{y} \pm t_{n_1+n_2-2,\frac{\alpha}{2}} S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ 2. $\left(\bar{x}-\bar{y} - t_{n_1+n_2-2,\alpha} S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \infty\right)$ 3. $\left(-\infty, \bar{x}-\bar{y} + t_{n_1+n_2-2,\alpha} S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$	1. $\bar{x}-\bar{y} \pm t_{v,\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ 2. $(\bar{x}-\bar{y} - t_{v,\alpha}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}, \infty)$ 3. $(-\infty, \bar{x}-\bar{y} + t_{v,\alpha}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}})$

**If σ^2 is known, then use z's instead of t's (see the case for known variance for one sample for details)

Power: Same as the one-sample case. For example, for the case (1) for equal variance assumption

$$\text{Power} = P_{\mu_1-\mu_2 \in H_a}\left(\left|\frac{(\bar{X}-\bar{Y}-\mu_0)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right| > t_{n_1+n_2-2,\frac{\alpha}{2}}\right)$$

Paired Samples: If the samples are paired, take the difference and use one-sample testing procedure. For example, $H_0: \delta = 0$ and $H_a: \delta \neq 0$ where δ represent the differences. Your n should be the number of pairs, S^2 should be the sample variance of the differences , and $\bar{X}$ is the mean of the differences