Last time: p-values

μ with $\overset{Z}{\text{known}}$ / $\overset{T}{\text{unknown}}$ σ.

$\mu_A - \mu_B$ " " Z " " if unknown, $\overset{T}{\sigma_A = \sigma_B}$, or n_A, n_B $\overset{Z}{\text{large}}$.

σ^2 $\quad \chi^2$

Ratio of variances

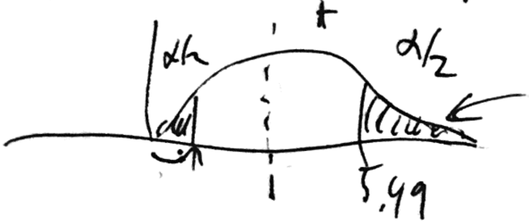
E.g.: Steel brands A & B. Compare variability.

$n_A = 13$, $S_A^2 = 19.2$, $n_B = 16$, $S_B^2 = 3.5$.

$H_0: \sigma_A^2 = \sigma_B^2 \quad H_1: \sigma_A^2 \neq \sigma_B^2$. $\quad C = \left\{ \frac{\sigma_A}{\sigma_B} > k \text{ or } \frac{\sigma_A}{\sigma_B} < \frac{1}{k} \right\}$

$$F = \frac{U/n}{V/m} = \frac{\cancel{(n_A-1)} S_A^2/\sigma_A^2 \,/\, \cancel{(n_A-1)}}{\cancel{(n_B-1)} S_B^2/\sigma_B^2 \,/\, \cancel{(n_B-1)}}.$$

Assume H_0: $\quad F = \frac{S_A^2}{S_B^2} = \frac{19.2}{3.5} = 5.49$



$$\text{p-val} = 2\left(1 - CDF\left(F_{12,15}, 5.49\right)\right).$$
$$= 0.0027.$$

If one sided $H_1: \sigma_A^2 > \sigma_B^2$, $\quad C = \left\{ \frac{\sigma_A^2}{\sigma_B^2} > k \right\}$.

Observe: $C : \left\{ \left| \log \frac{\sigma_A}{\sigma_B} \right| \geq k \right\}$. This critical region is equivalent to: $\log \frac{\sigma_A}{\sigma_B} > k$ or $\log \frac{\sigma_A}{\sigma_B} < -k$.

$$\frac{\sigma_A}{\sigma_B} > e^{k} \qquad\qquad \frac{\sigma_A}{\sigma_B} < e^{-k} = \frac{1}{e^{k}}.$$

P-values for proportions:

E.g.: 4 out of 20 patients suffered.
(Underlying assumptions: iid Bernoulli trials θ).
Test $H_0 : \theta = \underline{0.5}$ vs $H_1 : \theta \neq 0.5$.

RV: $n\bar{X} = \sum X_i \sim$ binomial w/ param $n=20, \theta=0.5$.

$$\mathbb{P}(n\bar{X} \leq 4) = \sum_{i=0}^{4} \binom{20}{i} \frac{(0.5)^{i}(0.5)^{20-i}}{(.5)^{20}} = 0.0059.$$

$n\bar{X} - 10 \leq -6 \qquad \mathbb{P}(|n\bar{X} - 10| \geq 6).$

p-val: 0.0118.

E.g. Oil company says $< 20\%$ all drivers have not used its oil. $H_0 : \theta \geq 20\% = 0.2$. $H_1 : \theta < 20\%$.
$\mathbb{P}(\text{not used oil}) = \theta$.

$H_0 : \theta \geq 20\% = 0.2$, $\underline{H_1 : \theta < .2}$.

Tested 200 drivers, of which 22 had not used their oil.

$n\bar{X}$ = binomial

$$\mathbb{P}(n\bar{x} \leq 22 ; \theta = .2) = \sum_{i=0}^{22} \binom{200}{i} .2^i (.8)^{200-i}$$
$$= 0.0005.$$

Alternatively, since $n = 200$ "large".

$$Z = \frac{n\bar{X} - n(.2)}{\sqrt{n(.2)(.8)}} = \frac{22 - 40}{\sqrt{200(0.2)(0.8)}} = -3.18.$$

$$\text{CDF} = \Phi(-3.18) = 0.0007.$$

<u>Differences among proportions</u>

Observed cell frequencies.

	# successes	# failures	total sample
Sample 1	f_{11}	f_{12}	$n_1 = f_{11} + f_{12}$.
Sample 2			n_2
Sample 3			⋮
⋮			
Sample k	f_{k1}	f_{k2}	n_k.

$k \times 2$ matrix of frequency values.

E.g.:

	Candidate A	Candidate B	totals
LA	232	168	400
NoLA	260	240	500
NY	197	203	400.

$H_0: \theta_1 = \theta_2 = \cdots = \theta_k = \theta_0$. (Think: Sample 1 = placebo).

If θ_0 known for some reason, use it.

If not (most of the time), need estimate.

$$\theta_0 = \hat{\theta} = \text{pooled estimate} = \frac{f_{11} + f_{21} + f_{31} + \cdots + f_{k1}}{n_1 + n_2 + \cdots + n_k}.$$

$$z_1 = \frac{n_1 \bar{X}_1 - n_1 \hat{\theta}}{\sqrt{n_1 \hat{\theta}(1-\hat{\theta})}}, \quad z_2 = \frac{n_2 \bar{X}_2 - n_2 \hat{\theta}}{\sqrt{n_2 \hat{\theta}(1-\hat{\theta})}} \cdots z_k.$$

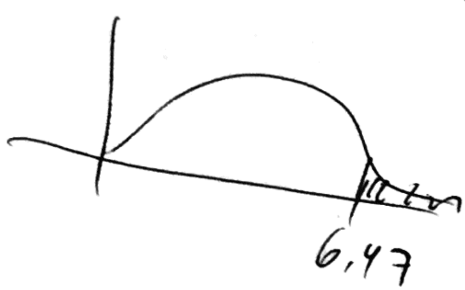
If we take $Z_1 + \dots + Z_k \sim N(0, k)$.

Bad idea: signs can cancel.

Good idea: take $Z_1^2 + Z_2^2 + \dots + Z_k^2 \sim \chi^2_{k-1}$

$$\chi^2_2 = \left(\frac{232 - 400(0.53)}{\sqrt{400(0.53)(0.47)}}\right)^2 + \frac{(260 - 500(0.53))^2}{500(0.53)(0.47)} + \left(\frac{(197 - 400(0.53))^2}{400(0.53)(0.47)}\right) = 6.47$$

$$\hat{\theta} = \frac{232 + 260 + 197}{1300} = 0.53.$$



p-val: $1 - \mathrm{CDF}(\chi^2_2, 6.47) = 0.039$.

$$\chi^2_{k-1} = \sum \frac{(f_{i1} - n_i \hat{\theta})^2}{\left(\sqrt{n_i \hat{\theta}(1-\hat{\theta})}\right)^2}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

expected cell values.

$e_{i1} = n_i \hat{\theta}$.

$e_{i2} = n_i (1 - \hat{\theta})$.