Lecture 24 (Ch. 9, 11)

We leaved Chi-squared test of k specific proportions in 1 pop.:

H_0: $\pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \cdots, \pi_k = \pi_{0k}$
H_1: At least 1 of These is wrong.
chi-sqd dist. with df $= k-1$

How does proportion of ($X=0,1$ something (e.g. tornadoes, $X=1$) vary across k categories (or k levels of a categorical var. Y)?
1 2-level X, 1 k-level Y.

And The chi-squared test of independence:

H_0: X and Y are indep.
H_1: X and Y are not indep.
chi-sqd dist. with df $= (k-1)(r-1)$

How does 1 categ./discrete var. (X) affect another categ./discrete var (Y)?
X, Y = 2 categ./discrete r.v.s.

This test is equivalent to a "test of homogeneity of population across categories" skipped skipped

In regression we studied How does 1 (or more) continuous var. (X) affect another continuous var. (Y)
X, Y = 2 continuous vars.

How about

How does 1 categ./discrete var. (X) affect 1 continuous r.v. (Y)?

This question requires comparing k means, ie.

μ_1 = mean of Y for $X=1$, μ_2 = mean of Y for $X=2$, ---- μ_k = ... $X=k$

And The question of whether X affects Y becomes

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ (Not $\mu_1 = \mu_{01}, \mu_2 = \mu_{02}, \ldots$)
H_1: At least 2 μ's are different.

Once again, This method compares The mean of a continuous r.v. at different levels of a categ./discr. r.v.

Example 1: Does knowledge of religion depend on religion?

ON RELIGION & PUBLIC LIFE | A project of the Pew Research Center

TOPICS
ISSUES
Abortion
Church-State Law
Death Penalty
Education
Gay Marriage & Homosexuality
Government
Politics & Elections
Science & Bioethics
Social Welfare
BELIEFS & PRACTICES
Belief in God
Frequency of Prayer
Importance of Religion
Religious Attendance
Other Beliefs & Practices
RELIGIOUS AFFILIATION
Christian
Jewish
Muslim
Other Affiliations
Unaffiliated
DEMOGRAPHICS
Age
Education & Income
Gender
Geography
Race
Other Demographics
REGIONS
Americas
Asia & the Pacific
Europe
Middle East & North Africa
Sub-Saharan Africa
PUBLICATIONS
Analyses
Event Transcripts
Graphics

pewforum.org › Topics › Beliefs & Practices

U.S. Religious Knowledge Survey

POLL • September 28, 2010

Executive Summary

Atheists and agnostics, Jews and Mormons are among the highest-scoring groups on a new survey of religious knowledge, outperforming evangelical Protestants, mainline Protestants and Catholics on questions about the core teachings, history and leading figures of major world religions.

On average, Americans correctly answer 16 of the 32 religious knowledge questions on the survey by the Pew Research Center's Forum on Religion & Public Life. Atheists and agnostics average 20.9 correct answers. Jews and Mormons do about as well, averaging 20.5 and 20.3 correct answers, respectively. Protestants as a whole average 16 correct answers; Catholics as a whole, 14.7. Atheists and agnostics, Jews and Mormons perform better than other groups on the survey even after controlling for differing levels of education.

Print Email Share AA Text Size

Atheists and Agnostics, Mormons and Jews Score Best on Religious Knowledge Survey

Average # of questions answered correctly out of 32

Total	**16.0**
Atheist/Agnostic	20.9
Jewish	20.5
Mormon	20.3
White evangelical Protestant	17.6
White Catholic	16.0
White mainline Protestant	15.8
Nothing in particular	15.2
Black Protestant	13.4
Hispanic Catholic	11.6

test scores (out of 32)

Looks like Athiests know most!?

In This Report

SCIENTIFIC AMERICAN™

Winner of the 2011 National Magazine Award for General Excellence

Search ScientificAmerican.com

News & Features | Blogs | Multimedia | Education | Citizen Science | Topics

Home » Scientific American Magazine » December 2010

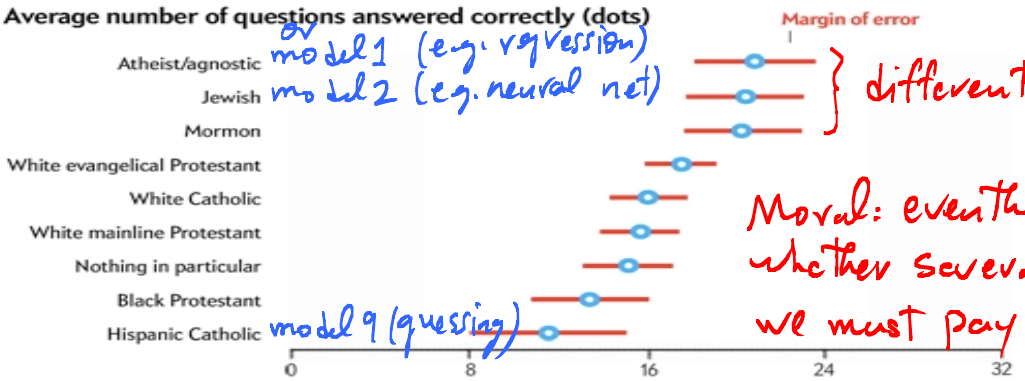
Advances | More Science

The Science of "Disestimation": The Shortcomings of Opinion Polls

Why we shouldn't put our faith in opinion polls

By Charles Seife | December 14, 2010 | 19

Share Email Print



or model 1 (e.g. regression)
model 2 (e.g. neural net)
} different?

model 9 (guessing)

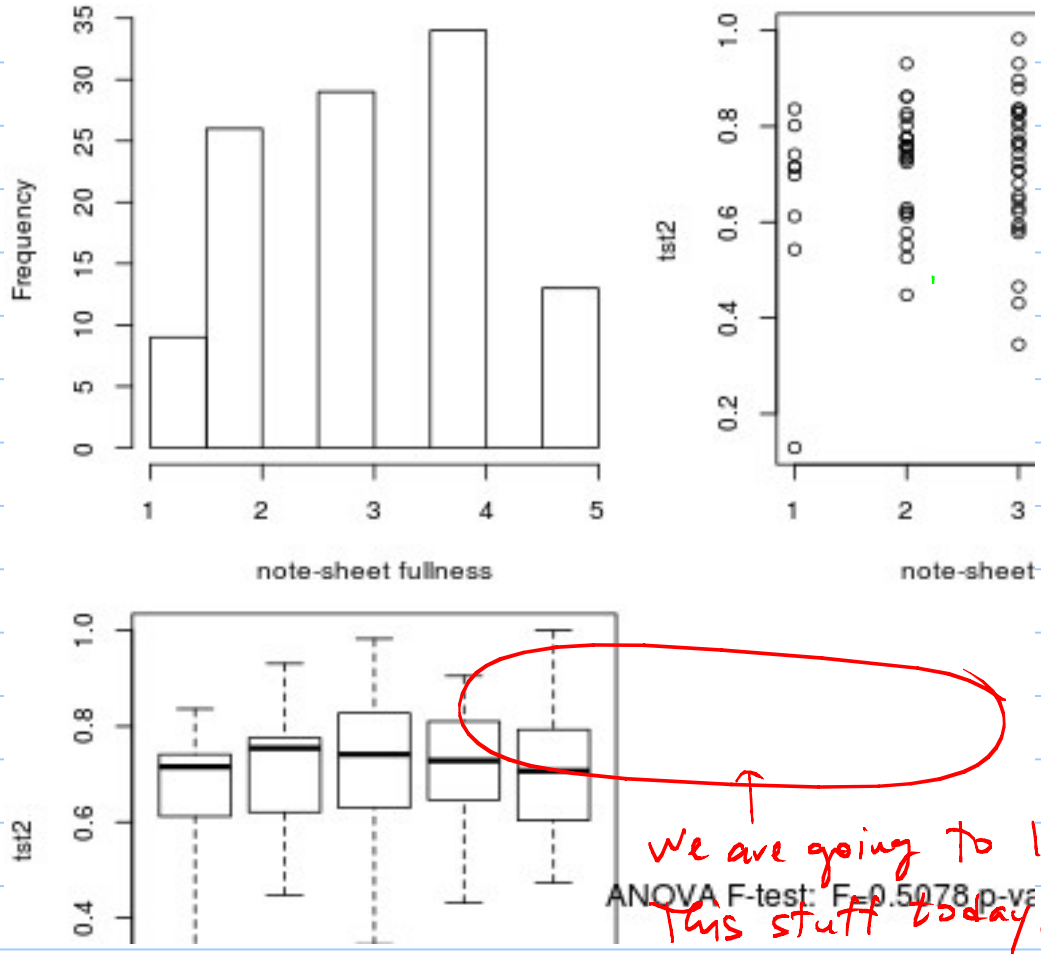
Moral: even though we are testing whether several means are equal, we must pay attention to variance!

Hence the name ANOVA

Example 2: Does fullness of note sheet have an effect on test scores?

note-sheet fullness		mean test score
not-so-full	1	0.6437
	2	0.7205
	3	0.7179
	4	0.7201
very full	5	0.7142

Again, looking at means is not enough. Must also look at variance.



We are going to learn this stuff today.

note that this test is just a generalization of the 2-sample/pop test (for comparing μ_1, μ_2) to the case of k populations.

Examples

Does mean knowledge of religion vary across different religions? ← above
Does mean test grade vary across fullness of cheat sheet? ← above
Does mean vibration vary across different types of ball bearing? ← example 9.1 in book
Does mean computer speed vary across different computers?
Does mean detection error vary across different detection algorithms?

Generally: Are k means different?

The Data will look like This

X (categ)

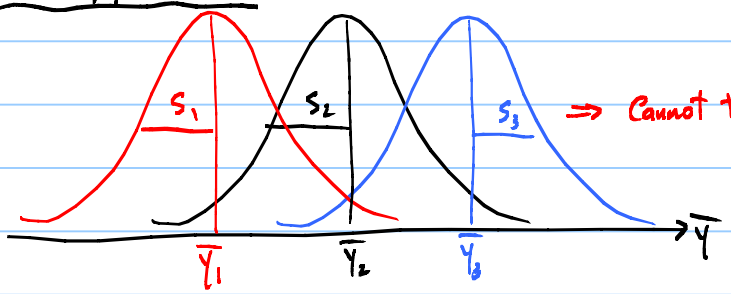
	1	2	3	4	5
Data on Y	⋮	⋮	⋮	⋮	⋮
	$\bar{y}_1 = \dots$	$\bar{y}_2 = \dots$	...	...	...
	$s_1 = \dots$	$s_2 = \dots$	...	...	...

Note That we are now dealing with a generalization of the 2-sample t-test to more than 2 pops, here, 5 populations.

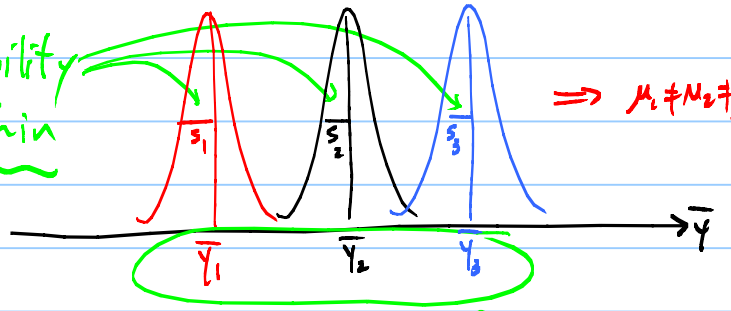
$H_0: \mu_1 = \mu_2 = \dots = \mu_5$

H_1: Atleast 2 μ's are diff.

The approach:



⇒ Cannot tell



Variability within

⇒ $\mu_1 \neq \mu_2 \neq \mu_3$

Variability between

The Way ANOVA answers That question is by finding out how much of The **total** variability in y is **within** each categ./pop., and how much is **between** The categ./pops.

This "within", "between" idea is a very powerful idea That shows-up everywhere!

Total variability in the y_{ij}

k categories/populations.

sample size in i^{th} pop.

$$n = \sum_{i=1}^{k} n_i$$

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$$

Grand mean

$$\frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) \bar{y}_i$$

j^{th} response in the i^{th} pop/Categ.

sample mean in i^{th} pop.

Sample var. in i^{th} pop.

$= (n_i - 1) s_i^2$

$$SST = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{\bar{y}})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$SS_{between\ group}$ = SS_{Tr} ← Treatment = $SS_{explained}$

Variation between groups ~ Sample var. of sample means

$SS_{within\ group}$ = SSE = $SS_{unexplained}$

Variation within groups ~ Sample mean of sample variances

FYI

If you see similarity with the ANOVA decomposition in regression, you are correct. There are similarities, and differences:

$$SS_{Total} = SS_{between} + SS_{within}$$

1-way ANOVA df: $n - 1 = (k - 1) + (n - k)$

k = # of levels in 1 factor (predictor) = # of pops.

Regression df: $n - 1 = k + [n - (k+1)]$

k = # of β's

different k's

Now, we can compare $SS_{between}$ and SS_{within}:

Note: If all the sample means are equal, then F=0.

Theorem:

If H_0 = True, $$F = \frac{SS_{between}/(k-1)}{SS_{within}/(n-k)} \equiv \frac{MS_{between}}{MS_{within}}$$

has an F-distribution with params. $df = (k-1, n-k)$

All we need is Table VIII

$$p\text{-value} = pr(F > F_{obs})$$

Again: The ANOVA F-test is a generalization of the 2-sample t-test to more than 2 populations.

FYI

One assumption of this Theorem is that the y's in each of the k populations are normal, and that they all have the same variance, ie. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$. (called homoscedasticity!)

Use qq plots to test this ↑ {See optional hw for understanding this}

example 9.1

Consider 5 brands of computers. A code has been run on each of the brands 6 times, and the completion times have been recorded. Here are the data:

Brand 1	2	3	4	5
13.1	⋮	⋮	⋮	⋮
15.0				
14.0				
⋮				
11.6				
Var. between ⇐ $\bar{y}_1 = 13.68$	$\bar{y}_2 = 15.97$	13.67	14.73	13.08
Var. within ⇐ $s_1 = 1.194$	$s_2 = 1.167$	---	---	---

$$\bar{\bar{y}} = \sum_{i=1}^{5} \left(\frac{n_i}{n}\right) \bar{y}_i = \frac{6}{30}(13.68) + \cdots = 14.22$$

$$SS_{between} = \sum_i n_i (\bar{y}_i - \bar{\bar{y}})^2 = 6(13.68 - 14.22)^2 + \cdots = 30.88$$

$$SS_{within} = \sum_i \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{5} (n_i - 1) s_i^2 = (6-1)(1.194)^2 + \cdots = 22.83$$

$$F_{obs} = \frac{30.88/(5-1)}{22.83/(30-5)} = 8.45$$

$df = (5-1, 30-5)$
See next page for how to read Table VIII.

p-value $= pr(F > F_{obs}) = pr(F > 8.45) < .001$

Conclusion: (at $\alpha = .01$) Reject H_0 ($\mu_1 = \mu_2 = \cdots$) in favor of H_1 (At least 2 μ's are diff)

which 2?
Section 9.3 skipped.

In English: Brand has an effect on speed.

Reading Table VIII

For the above example, we have $df_{num} = 4$, $df_{denom} = 25$

In Table VIII, for these df's you'll see

Area	F	
.1	2.18	→ if $F_{obs} = 2.18$, Then p-value = 0.1
.05	2.76	} → if $F_{obs} = 3$, Then .01 < p-value < .05
.01	4.18	
.001	6.49	} → if $F_{obs} = 8$, Then p-value < .001

FYI

Most software produce an ANOVA Table (See pre lab)

FYI

Source	df	SS	MS (mean)	F_{obs}	p-value
Between Group (factor)	k-1	$SS_{between}$			table VIII
Within Group (error)	n-k	SS_{within}			
Total	n-1	SS_{Total}			

from formula

For the above example (from R):

Source	df	SS	MS	F	p-value
factor	5-1	30.85	7.71	8.44	.00018
Error	30-5	22.84	0.91		
Total	30-1	53.7			

"large sample" → z
"small sample" → t

Summary:

z, t — $H_0: \mu = \mu_0$ $\quad H_1: \mu \square \mu_0$ — σ_x known / unknown

z — $H_0: \pi = \pi_0$ $\quad H_1: \pi \square \pi_0$ — z (No t). "large sample"

z, t — $H_0: \mu_1 - \mu_2 = \Delta_0$ $\quad H_1: \mu_1 - \mu_2 \square \Delta_0$ — indep. or paired

Chisqd — $H_0: \pi_1 = \pi_{01}, \pi_2 = \pi_{02}, \cdots \pi_k = \pi_{0k}$, H_1: Atleast 1 is wrong

chisqd — H_0: { 2 categ./discrete variables are independent H_1: not. / 2 pops are homogeneous w.r.t. k categories } (skipped.)

F — $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ $\quad H_1$: atleast 2 μ's are diff.

Again, note that the 1-way ANOVA problem deals with

$Y =$ { 1 cont. var. (whose mean we are interested in), and

$X =$ { 1 categ. var. (with k levels).

→ Next: Regression: 2 cont. variables, x, y.

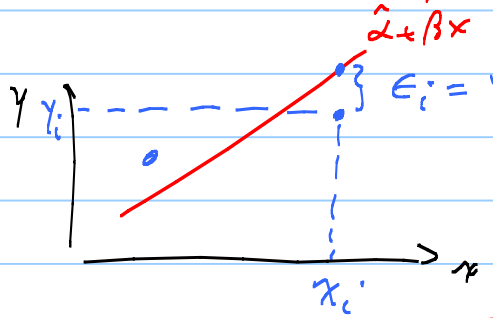
Ch. 11

We did regression $y_i = \alpha + \beta x_i + \cdots + \epsilon_i$ Ch. 3.

We did inference on μ, π, $\mu_1 - \mu_2$, $\pi_1 - \pi_2$, π_i, $\cdots$ $(\mu_1 = \mu_2 = \mu_3 = \cdots)$ Ch. 7, 8, 9

Now, we do inference in regression (on $\beta, \alpha, y(x)$, $\cdots$) Ch. 11.

Review:



arbitrary params to be estimated by OLS, ie. $\frac{\partial}{\partial \alpha} SSE$, etc.

⇒ For a sample we write $y_i = \alpha + \beta x_i + \epsilon_i$

and $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \Rightarrow \hat{y}(x) = \hat{\alpha} + \hat{\beta} x$

(a, b in book)

where $\hat{\alpha}, \hat{\beta}$ are the OLS estimates of α, β, ie.

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}.$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Recall that Sample var. $= s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{S_{xx}}{n-1}$

⇒ The Analysis of Variance:

$$SST = \sum (y_i - \bar{y})^2 = SS_{explained} + \underbrace{SS_{unexpl.}}_{SSE}$$

df: $\quad n-1 \quad = \quad k \quad + \quad n-(k+1)$

$R^2 = \frac{SS_{expl.}}{SST}$ — percent of variability in y explained by x ... (Goodness of fit)

k = # of β's (excluding α) = # of predictors in $y = \alpha + \beta_1 x_1 + \cdots \beta_k x_k$

$s_e = \sqrt{\frac{SSE}{n-(k+1)}}$ — std. dev. of errors ~ Typical error or spread about fit. (RMSE)

⇒ **For population**

Now, let's consider the population. For the moment suppose we have it. Just because we have the population, it does not follow that there is no scatter between x, y. I.e. even for the population, there is a scatter between x and y, and so there is an OLS fit for the pop!

⇒ I.e. even for the population there is an OLS fit! Call it the true fit. What symbols should we use to denote this true fit?

sample mean : $\bar{x}$	sample OLS fit : $\hat{\alpha}, \hat{\beta}$
true/pop./distr. mean : μ_x	true/pop./distr. OLS fit : ?

It's tempting to use α, β (w/o hat). But α, β are supposed to be free parameters that are tuned to minimize the SSE.
So, technically, we should introduce new symbols for the true OLS fit. However, to keep things simple, we will go ahead and use α, β to denote the true OLS fit.

⇒ I.e. up to now, α, β have been free parameters to do $\partial/\partial\alpha$, $\partial/\partial\beta$, ... But hence forth, α, β denote the OLS fits for the population.

⇒ In short: $\hat{y}(x) = \hat{\alpha} + \hat{\beta}x$ (for sample) and $y(x) = \alpha + \beta x$ (for population)

will be used for the respective predicted values.
I.e. $\hat{y}(x)$ = prediction (in sample)

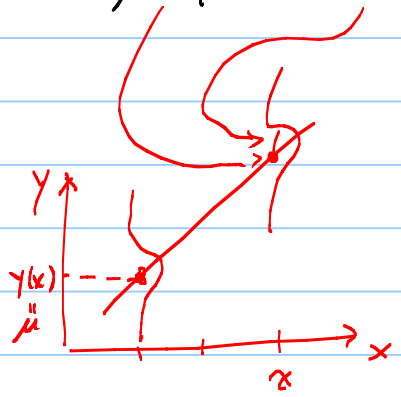
$y(x)$ = prediction (in pop.)

Now, to do inference we need a probability model (for regression):

Assume y's are Normally distr. at each x, with params $\mu = Y(x)$, $\sigma = \sigma_\epsilon$

e.g. $\mu = Y(x) = \alpha + \beta x + \dots$ $\quad$ $\sigma = \sigma_\epsilon$ = fixed

estimate α, β with $\hat\alpha, \hat\beta$ $\quad$ estimate with s_e



FYI At a given x, $y \sim N(Y(x), \sigma_\epsilon)$

$\therefore\ y = Y(x) + \epsilon \Rightarrow \epsilon = y - Y(x) \sim N(0, \sigma_\epsilon)$

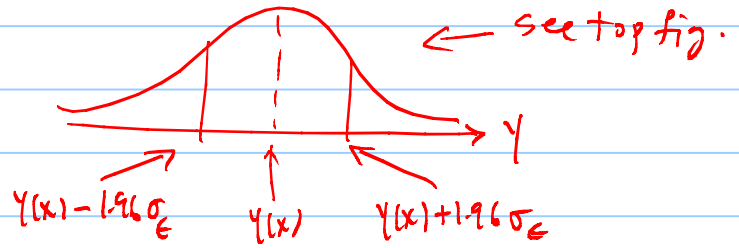
This allows us to say things like:

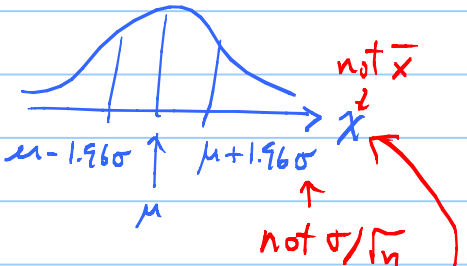
1) At a given x, $\hat y(x) = \hat\alpha + \hat\beta x$ estimates the true mean of y, $Y(x)$.

In short: For a fixed x, every thing we have done (CI, p-value...) now applies to y.

2) At a given x, we expect about 95% of the y's to be within $Y(x) \pm 1.96\sigma_\epsilon$.

like 95% of x's are within $\mu \pm 1.96\sigma$ (Ch.1)





3) At a given x, we can find other probs for y.

e.g. $\text{prob}(a < y < b \mid x) =$

True prediction = True mean | x.

$$\text{prob}\left(\frac{a - Y(x)}{\sigma_\epsilon} < \boxed{\frac{y - Y(x)}{\sigma_\epsilon}} < \frac{b - Y(x)}{\sigma_\epsilon}\right) = \text{Table I}$$

$z \sim N(0,1)$

like $\text{pr}(a < x < b) =$ (Ch.1)

$$\text{pr}\left(\frac{a-\mu}{\sigma} < \boxed{\frac{x-\mu}{\sigma}} < \frac{b-\mu}{\sigma}\right)$$

$z \sim N(0,1)$

Note that these are just prob calculations, not p-values or CIs.

hw_lect24_1:
The following data refer to the melting temperature, y (in some unit), of a certain material at four different pressures, x (in some unit).

Pressure	Temperature
1.6	59.5, 53.3, 56.8, 63.1, 58.7
3.8	55.2, 59.1, 52.8, 54.4
6.0	51.7, 48.4, 53.9, 49.0
10.2	44.6, 48.5, 41.0, 47.3, 46.1

a) Make a comparative boxplot of y for the four pressure levels. By R.

b) Based on the above boxplot, would you say there is a difference in the mean melting temperature for at least 2 of the pressure levels?

c) At alpha = 0.05, is there evidence that the mean melting temperature at the at least 2 of the four pressure levels are different? Report the p-value, and state the conclusion clearly. By R; see prelab to see how to do 1-way ANOVA in R.

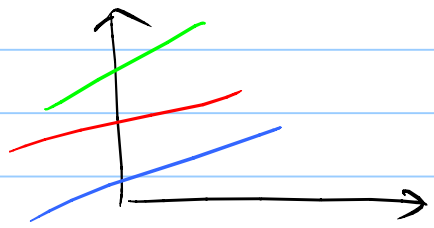
d) Write code to compute the above p-value "by hand," i.e. without using aov() or lm(), but using the basic formulas for SS_between, SS_within, etc.

e) After (or before) a 1-way ANOVA test, one should check the two assumptions that the y's are normally distributed within each group, and with the same variance. To that end, make a plot that shows four qqplots (one for each pressure level) superimposed onto a single figure; make sure that the four qqplots have different colors. Are the 4 qqplots reasonably straight, and do they have approximately equal slopes? Hint: in the first call to plot(), use xlim=c(-2,2) and ylim=range(y). Use the "By hand" method for making qqplots.

hw_lect24_2
In a problem dealing with flow rate (y) and pressure-drop (x) across filters, it is known that **y= -0.12 + 0.095 x**. Note: this is the true "fit" to the population. Suppose it is also known that sigma_epsilon = 0.025. Now, IF we were to make repeated observations of y when x=10, what's the prob. of a flow rate exceeding 0.835?

$y = -0.12 + 0.095x$, $\sigma_\varepsilon = 0.025$

hw_lect_optional: By hand or by R (see prelab)
Consider the data you collected. Take one of the continuous variables (call it y) and the categorical (or discrete) variable with 3 or more levels (call it x). Since x is discrete/categorical, we can consider each level as a different population. Eg, if your x has 3 levels (say H, M, L), separate the corresponding y's into 3 populations.
a) Do 1-way ANOVA to test if any of the k populations have different means. Report the p-value, and the conclusion In English.
b) Make qq-plots of y for each of the levels of X. It's important to have all qq-plots superimposed onto a single figure. See "By hand" q-plots in prelab. For example, if your x has 3 levels, then you need to have 3 qq-plots superimposed on one plot, e.g.



Recall that equal-slopes translate to equal variances, and so this will be a way of visually checking the homoscedasticity assumption mentioned above (in the lecture note).